

INTEGRATING SYNTAX, SEMANTICS, AND DISCOURSE  
DARPA NATURAL LANGUAGE UNDERSTANDING PROGRAM

R&D STATUS REPORT

**AD-A194 099** Unisys/Defense Systems

ARPA ORDER NUMBER: 5262

PROGRAM CODE NO. NR 049-602 dated 10 August 1984 (433)

CONTRACTOR: Unisys Defense Systems

CONTRACT AMOUNT: \$1,192,833

CONTRACT NO: N00014-85-C-0012

EFFECTIVE DATE OF CONTRACT: 4/29/85

EXPIRATION DATE OF CONTRACT: 4/28/89

PRINCIPAL INVESTIGATOR: Dr. Lynette Hirschman

PHONE NO. (215) 648-7554

SHORT TITLE OF WORK: DARPA Natural Language Understanding Program

REPORTING PERIOD: - 11/1/87-1/31/88

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

DTIC  
ELECTE  
APR 08 1988  
S D

DISTRIBUTION STATEMENT A  
Approved for public release  
Distribution Unlimited.

88 3 11 082

Partial contents!

## 1. Description of Progress

We are continuing to port Pundit to the Trident domain of maintenance reports.

A report was submitted to Darpa describing our research activities under the DARPA contract for the 18 month period from 6/86 to 12/87.

### 1.1. Grammar

The Trident domain provides a variety of complicated structures; thus work on Trident has led to improvements in the stable system grammar. There are also run-on sentences and fragments in the Trident messages; we have already extended PUNDIT's grammar to analyse these structures. In fact, this domain provides confirmation that the regular fragment types established in other domains occur in a wide variety of sublanguages. In the category of improvements to the stable system, we include the addition of rules and restrictions to permit multiple right modifiers of noun phrases, including appositives, as in *Replaced interlock switch with a new one (switch 7802) from supply*; a variety of fine-tunings to restrictions; and the addition of complex adjectival structures, which is in the final stages of implementation. These additions include equi structures such as *We are [unable to operate pump handle]* and raising structures such as *pump is [likely to operate]*; the ISRs developed for these two types of adjective complements will reflect their differing argument structure and referential possibilities.

### 1.2. Syntax/Semantics Interaction

We have finished the design of a more flexible interaction between syntax/semantics and are ready to start implementation. The disadvantages with the current selection mechanism are due mainly to the inflexibility of the patterns and the pattern matching. A bad pattern for the intransitive form of a verb such as *melt*, as in *flame melt*, cannot automatically be generalised to include the transitive form of the same verb, as in *X melted the flame*. In addition, the patterns are designed to rely on a straightforward match between the category of the selection restriction and the category of the head noun of the noun phrase. This makes it difficult to test more complicated semantic constraints such as *patient has symptom* corresponding to the sentence *Patient has stiff neck* (the computed attribute problem). The more flexible interaction we have in mind will consolidate the semantic information currently residing in the selection pattern database and the semantics rules file. This will immediately solve the generalisation problem, and will provide a much more powerful framework for exploring solutions to the computed attribute problem.

The implementation will replace the calls to VSO selection with calls to the semantic interpreter. This requires that a new mode of the semantic interpreter be defined which will test selection restrictions on verb arguments as quickly as possible. Our plan for achieving this speed up includes partially compiling the mapping process, as well as avoiding calls to pragmatics wherever possible. We will continue to use the stripper called by the current selection mechanism which simplifies the ISR, e.g., isolates head nouns of noun phrases, to produce a more canonical form for semantic testing. We will experiment with replacing the current mapping rules with "compiled" ISR skeletons, so that each different syntactic realisation of a particular verb will have a corresponding "stripped" ISR skeleton. Then the mapping process will be achieved by unifying the ISR skeleton with the "stripped" ISR for the actual sentence. These ISR skeletons will reside in a database, which the verb decompositions can index into. Since we will be avoiding calls to reference resolution, the semantic constraints (i.e., selection restriction) will have to be applied directly to the head nouns rather than the referents. The head nouns will be contained in the "stripped" ISR, and will unify with the logical variables in the ISR skeletons. These will be co-indexed with the appropriate logical variables for the thematic roles in the verb decompositions. So unifying the ISR skeleton with the sentence ISR will result in the thematic role variables being instantiated with the head nouns. Then each thematic role will have to be examined in turn, in order to apply the semantic constraint to the head noun. This will be done using the standard semantic analysis techniques, with a modified version of the "check" procedure which tests semantic constraints. We will compare processing times with the current mapping & checking process to determine time savings.

## 1.3. Semantics

The Trident verb decomposition rules have been extended so that Pundit's semantic interpreter can now process, at least in a rudimentary way, all seven of the messages in the test corpus.

Efforts are currently underway to fine-tune Pundit's performance in the Trident domain. Minor adjustments were required to allow the interpreter to provide appropriate representations for NQ structures—noun phrases consisting of a nominal followed by an alpha-numeric expression like *NOP 55a*. A more substantial change may be needed to provide a proper treatment of parenthetical expressions. Most of the parenthetical expressions in the Trident test corpus function as appositives. Thus, in the example below, the expressions *SW 7802* and *SNM 8026* are used to further describe what sort of individuals the referents introduced for *interlock switch* and *a new one* might be anchored to.

*Replaced interlock switch (SW 7802) with a new one (SNM 8026) from supply.*

The development of a treatment for parentheticals is being planned as part of a more general account of appositives.

## 1.4. Discourse

The Trident message-processing and database query application, in the domain of equipment failure messages, was found to differ from previous Pundit applications and domains in a number of interesting respects. In the course of addressing the theoretical and operational issues arising from this application, Pundit is being extended in directions which will enhance its capabilities as a general-purpose message processing system.

Key features of the application and domain which have required extensions to Pundit discourse-processing capabilities are as follows:

1. Messages: in previous domains, messages were simply blocks of text with no external structure. Trident messages, however, consist of five paragraphs, each of which addresses a specific topic (e.g. *first indication of trouble, part failure, probable cause*). Hence Trident messages are structured into multiple pre-defined segments.

The relationship between the heading of a segment and the segment itself bears a strong relationship to prompt-response pairs, which in turn resemble question-answer pairs. The responses can be direct (*probable cause: broken wire*), indirect (*the write head had black marks on it. This appears to be caused by...*), or meta-responses (*unknown*). In the case of indirect responses, the corresponding direct response must be inferred.

Analysis of messages revealed that there are interesting constraints on reference across segments. It appears that information from the formatted portions of messages is globally available to all segments, as is summary information from previously processed segments, but pronominal reference cannot access entities evoked from within a preceding segment.

2. Application: the application requires that Pundit process messages and provide a value for pre-specified database attributes for later queries. The crucial difference from previous applications is that the database relations and attributes are pre-defined, and in many cases there is not a direct correspondance between an item of data produced by Pundit's analysis and the database attribute to be valued. That is, the value of some attributes must be derived by inferencing. The progress made over the last quarter towards addressing these issues is described below.

1. A new message input front-end to Pundit has been developed which is capable of processing both formatted portions of messages and multiple segments, and which supports both batch and interactive modes. We anticipate generalising this approach across other domains, in such a way that front-ends can be readily tailored for each new application.

2. A prototype model of discourse structure was developed to capture constraints on reference across segments, and to facilitate database attribute valuation.

3. Analysis was begun for a potential redesign of Pundit's database update modules. It is anticipated that the new design will allow for predicate-driven mappings (where sets of semantic predicates map to database relations) as

well as attribute-driven mappings (where inferencing over the results of analysis is required to value the attribute).

4. Analysis was begun on a general approach to parsing and interpreting prompt-response pairs. Preliminary results indicate the need for a discourse-processor module, at a higher level than syntax and semantics: this module develops expectations about responses to prompts, communicates these expectations to syntax and semantics, validates the results against expectations, and in general exhibits a higher level of understanding than is possible with a purely sentence-based approach to discourse.

### 1.5. Demo Environment

Documentation on preparing and executing the demo for Pundit and its various domains has been compiled. It contains the most up-to-date information and directions on the subject matter and supercedes all previously made documentations and instructions. The development of an ISR pretty print procedure with Xwindow display facilities for the Pundit images on the Sun workstations has been initiated.

### 2. Change in Key Personnel

Korrinn Fu, who is receiving an M.S. in Computer Science from the Pennsylvania State University, started on November 16.

### 3. Summary of Substantive Information from Meetings and Conferences

#### 3.1. Darpa Meetings

Shirley Steele, Martha Palmer, and Lynette Hirschman attended the meeting of the Darpa Natural Language contractors at SRI.

Lynette Hirschman, Shirley Steele, and Deborah Dahl met in Cambridge on January 11-12 with MIT speech researchers Victor Zue and Stephanie Seneff on possible collaboration on a Spoken Language System.

#### 3.2. Papers and Presentations

Martha Palmer presented an invited talk at Bell Labs entitled, "Developing and Porting a Text Processor", December 11.

A paper describing the treatment of fragments in the PUNDIT system has been accepted for presentation at the annual meeting of the Association for Computational Linguistics.

A draft of a paper describing our experience with Pundit as a large Prolog program has been prepared for submission to the 1988 Joint Conference on Logic Programming and is being reviewed.

A paper authored by Shirley Steele of PRC and Richard Sproat of Bell Labs was presented at the November meeting of the Acoustical Society of America. A paper by Steele and Janet Pierrehumbert of Bell Labs was accepted for publication in *Phonetica*.

### 4. Problems Expected or Anticipated

None.

### 5. Action Required by the Government

None.



**6. Fiscal Status**

- (1) Amount currently provided on contract:  
     \$ 1,192,833 (funded)                      \$1,704,901 (contract value)
- (2) Expenditures and commitments to date:  
     \$ 889,345
- (3) Funds required to complete work:  
     \$ 815,556

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>per ltr.</i>	
Distribution	
Availability Codes	
Dist	Avail and/or Special
A-1	

